STRUCTURE OF RANDOM r-SAT BELOW THE PURE LITERAL THRESHOLD

ALEXANDER D. SCOTT* AND GREGORY B. SORKIN

Abstract. It is well known that there is a sharp density threshold for a random r-SAT formula to be satisfiable, and a similar, smaller, threshold for it to be satisfied by the pure literal rule. Also, above the satisfiability threshold, where a random formula is with high probability (**whp**) unsatisfiable, the unsatisfiability is **whp** due to a large "minimal unsatisfiable subformula" (MUF).

By contrast, we show that for the (rare) unsatisfiable formulae below the pure literal threshold, the unsatisfiability is **whp** due to a unique MUF with smallest possible "excess", failing this **whp** due to a unique MUF with the next larger excess, and so forth. In the same regime, we give a precise asymptotic expansion for the probability that a formula is unsatisfiable, and efficient algorithms for satisfying a formula or proving its unsatisfiability. It remains open what happens between the pure literal threshold and the satisfiability threshold. We prove analogous results for the k-core and k-colorability thresholds for a random graph, or more generally a random r-uniform hypergraph.

1. Introduction

Let $r \geq 3$, and consider a random r-SAT formula F with n variables, where each of the $2^r\binom{n}{r}$ possible clauses is present independently with probability $p = \alpha n^{-(r-1)}$. Friedgut [10] showed that there is a threshold $c_r = c_r(n)$ for satisfiability: for every $\varepsilon > 0$, as $n \to \infty$, if $\alpha < (1-\varepsilon)c_r$ then F is with high probability (**whp**, i.e., asymptotically almost surely) satisfiable, while if $\alpha > (1+\varepsilon)c_r$ then F is **whp** unsatisfiable. For unsatisfiable formulae, it is natural (and useful) to ask why. If F is unsatisfiable then it has one or more minimal unsatisfiable subformulae (MUFs); these are the minimal "obstacles" to satisfiability. Chvátal and Szemerédi [5] showed that, in the unsatisfiable regime (up to very high clause density) a random formula will not contain any small unsatisfiable subformula. Thus such a formula is typically unsatisfiable for a non-local reason, which also makes it difficult to prove unsatisfiability.

The aim of this paper is to develop an analogous picture for the rare unsatisfiable r-SAT formulae *below* the satisfiability threshold, and to investigate its algorithmic consequences. We are unable to completely characterize unsatisfiable formulae below the satisfiability threshold c_r, but we can do so below the smaller "pure literal" threshold α_r^*. We show that such a formula F is typically unsatisfiable for a *small* reason. Specifically, ranking MUFs in terms of *excess* ($r-1$ times the number of clauses, less the number of variables) only certain excesses are possible, and there are only finitely many MUFs with any given excess. Theorem 10 asserts that, **whp**, F contains a unique MUF, and this MUF has the minimum possible excess. Furthermore, if we condition on F having no MUF with excess up to i, then **whp** F still contains a unique MUF, and this MUF has the minimum possible excess greater than i. Additionally, Theorem 12 gives a precise asymptotic expansion for the probability of unsatisfiability: it is a power series in $1/n$, each of whose coefficients is an explicitly computable polynomial evaluated at α. (Failure of the pure literal rule, in place of unsatisfiability, is characterized similarly, but in terms of minimal full formulae, MFFs.)

* This research was supported in part by EPSRC grant GR/S26323/01, and by DIMACS, Center for Discrete Mathematics and Theoretical Computer Science, Rutgers, the State University of New Jersey, funded by the National Science Foundaton under Grant No. DMS06-02942, Special Focus on Discrete Random Systems.

We also consider failure of the pure literal rule (in place of unsatisfiability), obtaining a similar characterization, but in terms of minimal full subformulae (in place of minimal unsatisfiable subformulae), and a similar asymptotic expansion for the probability that the pure literal rule fails.

For random graphs and r-uniform hypergraphs (in place of r-SAT formulae), we develop a completely analogous picture for k-colorability and the existence of a nonempty k-core (in place of satisfiability and failure of the pure literal rule, respectively).

Algorithmically, our results immediately imply that for a typical unsatisfiable formula in the pure literal regime (a typical atypical formula), we can quickly find a witness. Additionally, we show that for sufficiently sparse random formulae (possibly below the pure literal threshold), in polynomial expected time we can decide satisfiability, output a satisfying assignment for satisfiable formulae, and for unsatisfiable formulae, output both an assignment satisfying as many clauses as possible, and a minimal unsatisfiable subformula (with corresponding results for hypergraphs). The hope is for algorithms efficient up to the pure literal threshold, and if possible up to the satisfiability threshold. (That goal was already achieved for the special case of 2-variable clauses, namely the class Max 2-CSP encompassing Max Cut, Max 2-SAT, the Ising model, and more. There, the two thresholds coincide, and [20] gave an algorithm running in expected linear time, exploiting the exponentially small probability of components of large excess.)

Stepping back, our exploration of unsatisfiable formulae in the satisfiable regime is complementary to existing explorations of the other three cases. Characterization of unsatisfiable formulae in the unsatisfiable regime was the main goal of [5]. Algorithms for satisfiable formulae in the unsatisfiable regime are often sought in the "planted" model, but recently there has been success in the uniform model [8]. Vast attention has been paid to algorithms for satisfiable formulae in the satisfiable regime, and we note just one recent result, [7].

A similar type of structural result — where if a likely property fails to hold, it most likely does so for a smallest reason, otherwise most likely for a second-smallest reason, and so forth — occurs in the context of random triangle-free graphs, although the proofs are completely different. A random triangle-free graph is **whp** bipartite [9], and otherwise can **whp** be made bipartite by deleting one vertex, otherwise **whp** by deleting two vertices, and so on [18]. It would be interesting to see other examples of this phenomenon.

2. Structural results for random instances of r-SAT

In this section, we prove our results for random instances of r-SAT. In order to prove our main result, we must first build up a structural picture of random formulae. Any minimum unsatisfiable formula must be *full* (all variables appear both with and without negation), and it turns out to be simpler to concentrate on full subformulae rather than minimum unsatisfiable subformulae. We divide our analysis into three ranges:

- *Subformulae of size at most K:* In this range, we determine rather precisely the joint distribution of full subformulae.
- *Subformulae of size between K and εn:* We show that with probability $O(n^{-s})$ there are no full subformulae in this range.
- *Subformulae of size at least εn*: We show that, with exponentially small failure probability, there are no full subformulae in this range (provided the density is below the pure literal threshold).

Here, we can choose any value for s, and then K and $\varepsilon > 0$ are carefully chosen constants (K must be sufficiently large in terms of s, and then ε must be sufficiently small in terms of K), while n is the number of variables. We begin in Section 2.1 by giving definitions. The analysis for the three ranges is given in Sections 2.2, 2.3 and 2.4; we put the pieces together in Section 2.5.

2.1. Basic definitions and random model.

A *conjunctive normal form* (CNF, or "SAT") formula consists of a set of *literals* (signed *variables*, i.e., variables and their negations) and a set of

clauses over these literals, each clause comprised of distinct variables with arbitrary signs. In an r-SAT formula each clause contains r literals; note that for a formula on n variables there are $2^r\binom{n}{r}$ possible r-clauses. A formula F is *satisfiable* if there is some assignment of True and False values to its variables such that each clause contains at least one True literal (a literal corresponding to a variable inherits its truth assignment, while the negated variable gets the negated assignment).

We define a *random formula* $F \in \mathcal{F}^r_{n,p}$ in analogy with a random graph $G \in \mathcal{G}_{n,p}$, letting each possible r-clause be present with probability p. We are primarily interested in random formulae where the expected number of clauses scales linearly with the number of variables. In any case, we work with three parametrizations, given by p, c, and α (all potentially functions of n), related by

$$p = \frac{cn}{2^r\binom{n}{r}} = \alpha n^{-(r-1)}, \tag{1}$$

where p is the clause probability, cn is the expected number of clauses, and α is a parametrization that is convenient because it is in fixed proportion to p but has the same desirable scaling behavior as c, since $\alpha = (1 + O(1/n))2^{-r}r!c$.

The *order* $|H|$ of a formula H is the number of variables (not literals); the *size* $e(H)$ is the number of clauses. We call a formula *empty* if it has no clauses, i.e., $e(H) = 0$. We define the *excess* of a formula in analogy with an established definition for hypergraphs, itself a natural extension of the excess (of edges over vertices) of a graph:

$$\operatorname{ex}(H) = (r-1)e(H) - |H|. \tag{2}$$

Two order-n formulae H and H' are isomorphic if there is remapping of their variables and their signs (under the action of the obvious group with $2^n n!$ elements). An *automorphism* of H is an isomorphism between H and itself, and we write $\operatorname{aut} H$ for the automorphism group.

H is a (proper) *subformula* of F if H's variable and clause sets are subsets of F's (and at least one of the containments is proper). We shall say that H' is a *copy* of H in F if H' is a subformula of F that is isomorphic to H (note that the isomorphism might involve changing signs). If F has any subformula H' isomorphic to H we may simply say that F *contains* H.

For formulae H and F, we write $X_H(F)$ for the number of copies of H in F. For a random formula $F \in \mathcal{F}^r_{n,p}$, recalling (1) and (2) and using the falling factorial notation $n_{(k)} = n(n-1)\cdots(n-k+1)$,

$$\begin{aligned}
\mathbb{E}X_H &= \frac{1}{|\operatorname{aut} H|}\binom{n}{|H|}|H|!2^{|H|}p^{e(H)} \\
&= \tfrac{1}{|\operatorname{aut} H|}n_{(|H|)}2^{|H|}\left(\alpha n^{-(r-1)}\right)^{e(H)} \\
&= \frac{n_{(|H|)}}{n^H}\frac{2^{|H|}}{|\operatorname{aut} H|}\alpha^{e(H)}n^{-\operatorname{ex}(H)} && (3) \\
&= (1+O(1/n))\tfrac{2^{|H|}}{|\operatorname{aut} H|}\alpha^{e(H)}n^{-\operatorname{ex}(H)}. && (4)
\end{aligned}$$

We say that a literal of F is *pure* if its complement does not appear in any clause of F. The *pure literal rule* chooses a pure literal of F (if there is any), and produces a smaller formula F' by deleting the literal's variable from F's set of variables, and deleting all clauses containing the literal from F's set of clauses. Note that F is satisfiable iff F' is, and if F is satisfiable then a satisfying assignment for F can be recovered from a satisfying assignment to F' by setting the selected literal True. The pure literal rule succeeds if F is eventually reduced to an empty formula, for then it produces a satisfying assignment for F; otherwise it is said to fail (and no conclusion can be drawn about the satisfiability of the original formula).

We call a formula H *full* if it is nonempty and has no pure literals (i.e., every variable and complemented variable of H appears in some clause); we say that H is a *full formula* (FF). We call a formula H a *minimal full formula* (MFF), if H is full and has no full proper subformula. It

is well known, and easy to see, that, regardless of how the pure literal rule chooses pure literals, it fails on F iff F contains a full subformula or equivalently iff F contains a MFF.

We call a formula H a *minimal unsatisfiable formula* (MUF) if H is unsatisfiable and contains no unsatisfiable proper subformula. It is clear that F is unsatisfiable iff it contains a MUF (F may itself be a MUF, or may properly contain one or more MUFs), and that a MUF is necessarily a FF. For a formula F, a contained MUF can be thought of as an obstruction to F's satisfiability, and a contained MFF as an obstruction to satisfying F using the pure literal rule. We will be interested in the probability that a random formula contains MUFs and MFFs of various sizes, and in particular whether typical obstructions are large or small.

2.2. **Small subformulae.** We begin by considering subformulae of constant size, and give fairly precise results for their distribution. These results hold for random formulae of any bounded density $c = c(n) = O(1)$ (equivalently $\alpha = \alpha(n) = O(1)$).

Lemma 1. *Suppose that $r \geq 3$. If H is full then $\mathrm{ex}(H) > 0$. Furthermore, for every $s > 0$, there are (up to isomorphism) only finitely many full formulae H with $\mathrm{ex}(H) = s$.*

Proof. If H is a full formula of order t, then by definition each of the t variables of H must occur at least twice (once with each sign) in the clauses of H. So $e(H) \geq 2|H|/r$, which implies

$$\mathrm{ex}(H) \geq 2(r-1)|H|/r - |H| = (r-2)|H|/r.$$

Since $r > 2$, this is strictly positive, the lemma's first assertion. Flipping the inequality, if $\mathrm{ex}(H) = s$ then $|H| \leq rs/(r-2)$, which implies that there are only finitely many possibilities for H. □

Since every MUF is a FF, there are also finitely many MUFs of each excess.

The following proposition shows that *fullness* plays a role somewhat like that of *strict balance* condition for graphs (see for example [2, Chapter IV]). A strictly balanced graph is one where every proper subgraph has strictly smaller density (ratio of edges to potential edges), and this can be used to show that a union of two strictly balanced graphs of equal density is a graph with strictly greater density. Here we have a property of a stronger type: the union of two non-nested full formulae (with possibly different excesses) is a formula with excess strictly greater than that of either.

Proposition 2. *Suppose that $r > 2$. For full formulae H_1 and H_2, with $H_1 \not\subseteq H_2$, $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + 1$.*

Proof. If $V(H_1) \subseteq V(H_2)$ then $|H_1 \cup H_2| = |H_2|$ while $e(H_1 \cup H_2) > e(H_2)$, implying $\mathrm{ex}(H_1 \cup H_2) > \mathrm{ex}(H_2)$. Since ex is integer-valued, this implies $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + 1$.

Otherwise, let $t = |V(H_1) \setminus V(H_2)| > 0$. Then $H_1 \cup H_2$ contains $2t$ more literals than H_2, and therefore contains at least $2t/r$ more clauses. So $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + (r-1)2t/r - t = \mathrm{ex}(H_2) + \frac{r-2}{r}t > \mathrm{ex}(H_2)$. Since ex is integer-valued, this implies $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + 1$. □

Claim 3. *Let $r \geq 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. For any fixed, full formula H,*

$$\mathbb{P}(\exists \text{ a copy of } H \text{ in } F) = (1 + O(1/n))\tfrac{2^{|H|}}{|\operatorname{aut} H|}\alpha^{e(H)}n^{-\mathrm{ex}(H)}.$$

Proof. With X_H the number of copies of H in F, the probability in question is $\mathbb{P}(\exists \text{ a copy of } H \text{ in } F) = \mathbb{P}(X_H > 0)$. It follows from inclusion-exclusion that

$$\mathbb{E}X_H \geq \mathbb{P}(X_H > 0) \geq \mathbb{E}X_H - \frac{1}{2}\mathbb{E}X_H(X_H - 1). \tag{5}$$

We will exploit Proposition 2 to show that $\mathbb{E}X_H(X_H - 1)$ is small compared with $\mathbb{E}X_H$.

We know already from (4) that

$$\mathbb{E}X_H = (1 + O(1/n))\tfrac{2^{|H|}}{|\operatorname{aut} H|}\alpha^{e(H)}n^{-\mathrm{ex}(H)}. \tag{6}$$

Note that $X_H(X_H-1)$ is the number of ordered pairs $\langle H_1, H_2\rangle$ of distinct (but possibly overlapping) copies of H in F. Let $\mathcal{H}$ be the set of isomorphism classes of *all* formulae $H' = H_1 \cup H_2$ with H_1 and H_2 isomorphic to H. Note that $\mathcal{H}$ is a finite collection of formulae and depends on H alone, not F or n: to enumerate $\mathcal{H}$ it suffices to consider formulae H_1 and H_2 on variables $1,\dots,2|H|$. Each copy in F of $\langle H_1, H_2\rangle$, corresponds in a 1-to-1 fashion to a copy in F of some $H' \in \mathcal{H}$ along with a covering of H' by an ordered pair $\langle H_1, H_2\rangle$ where H_1 and H_2 are both subformulae of H' and are both isomorphic to H. For $H' \in \mathcal{H}$, let $b(H')$ denote the number of ways of writing H' as a union of an ordered pair $\langle H_1, H_2\rangle$ of subformulae of H' that are copies of H. Then we have

$$\begin{aligned}
\mathbb{E}[X_H(X_H-1)] &= \sum_{H'\in\mathcal{H}} b(H')\mathbb{E}(X_{H'}(F)) \\
&= (1+O(1/n)) \sum_{H'\in\mathcal{H}} b(H')\tfrac{2^{|H'|}}{|\operatorname{aut} H'|}\alpha^{e(H')}n^{-\operatorname{ex}(H')} \text{ (by (4))} \\
&\le (1+O(1/n)) \left(\sum_{H'\in\mathcal{H}} b(H')\tfrac{2^{|H'|}}{|\operatorname{aut} H'|}\right)\alpha^{e(H)+1}n^{-(\operatorname{ex}(H)+1)} \\
&= O(1)\ \alpha^{e(H)+1}n^{-(\operatorname{ex}(H)+1)} \\
&= O(\alpha/n)\ \mathbb{E}[X_H],
\end{aligned}$$

where the inequality uses Proposition 2, the following equality uses that the set $\mathcal{H}$ is independent of F, and the final line similarly uses that the $\frac{2^{|H|}}{|\operatorname{aut} H|}$ in $\mathbb{E}[X_H]$ (see (4) again) is independent of F, and $\alpha = O(1)$.

With (5) and (6) this establishes the claim. □

Claim 3 already tells us something about the likelihood of small subformulae. Medium and large subformulae will be treated in subsequent sections, but while we are considering fixed subformulae we give two more lemmas that will be used for the structural results of Theorems 10 and 11.

Lemma 4. *Let $r \ge 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. Let H_1 and H_2 be fixed full formulae. Then*

$$\mathbb{P}(F \text{ contains non-nested copies of } H_1 \text{ and } H_2) = O\Big(n^{-\max\{\operatorname{ex}(H_1),\operatorname{ex}(H_2)\}-1}\Big).$$

Proof. Let $\mathcal{H}$ be the set of all isomorphism classes of unions of a copy of H_1 and a copy of H_2, where the two copies are not nested. By Proposition 2, any $H' \in \mathcal{H}$ has $\operatorname{ex}(H') \ge \max\{\operatorname{ex}(H_1), \operatorname{ex}(H_2)\}+1$ and so the assertion follows from Claim 3 by summing over $\mathcal{H}$. (As in the previous proof, $\mathcal{H}$ is a finite set, and is independent of F and n.) □

Lemma 5. *Let $r \ge 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. If $H_1,\dots,H_s$ are distinct FFs then*

$$\mathbb{P}(F \supset H_1 \mid F \not\supset H_2,\dots,F \not\supset H_s) = (1+O(1/n))\mathbb{P}(F \supset H_1).$$

Proof. First consider the case of just two FFs. Because H_1 and H_2 are distinct, they cannot be nested, and so we can use Lemma 4. Let E_i be the event that F contains a copy of H_i. Then

$$\mathbb{P}(E_1 \mid \neg E_2) = \frac{\mathbb{P}(E_1 \cap \neg E_2)}{\mathbb{P}(\neg E_2)} = \frac{\mathbb{P}(E_1) - \mathbb{P}(E_1 \cap E_2)}{1-\mathbb{P}(E_2)} = (1+O(1/n))\mathbb{P}(E_1),$$

where the last equality follows from Claim 3 and Lemma 4.

In the general case,

$$\mathbb{P}(\bigcap_{i=2}^{k} \neg E_i) \ge 1 - \sum_{i=2}^{k} \mathbb{P}(E_i) = 1 - O(1/n).$$

Also,

$$\mathbb{P}(E_1 \cap \bigcap_{i=2}^{k} \neg E_i) \geq \mathbb{P}(E_1) - \sum_{i=2}^{k} \mathbb{P}(E_1 \cap E_i) = \mathbb{P}(E_1) - O(\mathbb{P}(E_1)/n),$$

where the last equality follows from Claim 3 and Lemma 4. Combining,

$$\mathbb{P}(E_1 \mid \bigcap_{i=2}^{k} \neg E_i) = \frac{\mathbb{P}(E_1 \cap \bigcap_{i=2}^{k} \neg E_i)}{\mathbb{P}(\bigcap_{i=2}^{k} \neg E_i)} = (1 + O(1/n))\mathbb{P}(E_1).$$

□

2.3. **Medium subformulae.** We now turn to a middle range of subformula size, namely between a large constant and a small linear size. Once again, our results hold at all densities with α bounded.

The following is the sort of bound computed in [5].

Lemma 6. *Let $r \geq 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n)$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. For $1 \leq t \leq n/2\alpha^{1/(r-1)}$, the probability that F contains any full subformula with t variables is at most*

$$\left(\left(4^{(r-1)/r} e \alpha^{2/r}\right)(t/n)^{1-2/r}\right)^t. \tag{7}$$

Proof. Let the set of variables be $v_1, \ldots, v_n$. We order all $2n$ literals as $v_1 < \neg v_1 < v_2 < \neg v_2 < \cdots$.

A full subformula H of F with order t must contain at least $2t/r$ clauses. We let $s = \lceil 2t/r \rceil$ and define a subformula $H^* = H^*(H)$ with s clauses as follows. Let L be the set of $2t$ literals occurring in clauses of H. Let x_1 be the smallest literal in L, and let C_1 be the lexicographically smallest clause of H (sorting the literals within each clause as above) that contains x_1. For $i = 2, \ldots, s$, let x_i be the smallest literal in L that does not appear in any C_j, $j < i$, and let C_i be the lexicographically smallest clause of H that contains x_i. (x_i is well defined since we are always excluding literals from at most $s-1$ clauses, which together contain at most $(s-1)r < 2t$ distinct literals.) We then take H^* to be the conjunction of $C_1, \ldots, C_s$.

Over *all* full formulae H on a given set of t variables, the number of formulae $H^* = H^*(H)$ is at most $\binom{2t}{r-1}^s$ (there are at most $\binom{2t}{r-1}$ choices for each C_i, as it is forced to contain x_i), so the number of formulae of type H^* that could possibly be subformulae of F is at most $\binom{n}{t}\binom{2t}{r-1}^s$. Let X be the number of full subformulae of F with order t, and let Y be the number of subformulae of type H^* of F. Then clearly $X > 0$ implies $Y > 0$ (if X counts H, then Y counts $H^*(H)$), so

$$\begin{aligned}
\mathbb{P}(X > 0) \leq \mathbb{P}(Y > 0) \leq \mathbb{E}(Y) &\leq \tbinom{n}{t}\tbinom{2t}{r-1}^s p^s \\
&\leq (en/t)^t (2t)^{s(r-1)} (\alpha/n^{r-1})^s \\
&= (en/t)^t \left(\alpha (2t/n)^{(r-1)}\right)^s \\
&\leq (en/t)^t \left(\alpha (2t/n)^{(r-1)}\right)^{2t/r},
\end{aligned}$$

which equals (7). □

Corollary 7. *Let $r \geq 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. For any positive integer s, there exist an integer $t_0 > 0$ and a real value $\varepsilon_0 > 0$ such that the probability that F contains any full subformula with between t_0 and $\varepsilon_0 n$ variables is $o(n^{-s})$.*

Proof. Since the probability above is increasing in α, it is enough to prove the result for α constant, replacing $\alpha(n)$ by $\alpha = \max\{\sup_n \alpha(n), 1\}$. We first choose ε_0 small enough that $\varepsilon_0 < 1/2\alpha^{1/(r-1)}$ (so that any $t \leq \varepsilon_0 n$ satisfies the hypothesis of Lemma 6) and that

$$4^{(r-1)/r} e \alpha^{2/r} \varepsilon_0^{1-2/r} \leq 1/e.$$

Thus (7) is at most e^{-t} for all $0 < t \le \varepsilon_0 n$. Summing over t, it follows that the probability that F contains a full subformula with between $2s \log n$ and $\varepsilon_0 n$ variables is $o(n^{-s})$.

Now let $t_0 = 1 + \lceil sr/(r-2) \rceil$. For $t_0 \le t \le 2s \log n$, (7) is at most

$$\left(4^{(r-1)/r} e \alpha^{2/r} (2s \log n/n)^{1-2/r}\right)^{t_0} \le \left(\frac{8e\alpha s \log n}{n^{(r-2)/r}}\right)^{t_0} = O(n^{-s-1/r}(\log n)^{s+1}) = o\left(\frac{n^{-s}}{\log n}\right).$$

So the probability that F contains a full subformula with between t_0 and $2s \log n$ variables is $o(n^{-s})$. □

2.4. **Large subformulae.** Finally, we show that large subformulae are unlikely. This is the most delicate regime, and we will need to bound α more strictly. Some bound on α is certainly necessary: if α lies above the *satisfiability* threshold then a random subinstance is **whp** unsatisfiable, but (as shown by Chvátal and Szemerédi [5]) **whp** any unsatisfiable subinstance has size $\Omega(n)$. We will prove that large subformulae are unlikely for α below the *pure literal* threshold; what happens between the two thresholds is an open question.

Molloy [17] showed that there is a sharp threshold for the pure literal rule. Specifically, for $r \ge 3$, the threshold is[1]

$$\alpha^* = \min_{y>0} \frac{(r-1)!y}{2^{r-1}(1-e^{-y})^{r-1}}. \tag{8}$$

For any constant α, letting $p = \alpha n^{-(r-1)}$ and letting $F \in \mathcal{F}^r_{n,p}$ be a random formula,

$$\mathbb{P}(\text{pure literal rule finds a solution}) \to \begin{cases} 1 & \text{if } \alpha < \alpha^* \\ 0 & \text{if } \alpha > \alpha^*. \end{cases}$$

Achlioptas and Peres showed [1] that, as $r \to \infty$, the threshold for satisfiability (though not proved to be a constant rather than a function of n) is $c_{\mathrm{SAT}} = (1 + o(1))2^r \log 2$, leading via (1) to $\alpha_{\mathrm{SAT}} = (1 + o(1))r! \log 2$. By setting $y = r$ in (8) one can verify that the thresholds α^* and α_{SAT} diverge for large r: the gap in our knowledge of the behavior between the two is a wide one.

We need to show that large minimal unsatisfiable subinstances are unlikely; we therefore need a large deviation bound for values of α below the satisfiability threshold. We shall need the following version of the Azuma-Hoeffding inequality, given by McDiarmid [12].

Lemma 8. *Let $X_1, \dots, X_n$ be independent random variables, with X_k taking values in a set A_k for each k. Suppose that a measurable function $f : \prod A_k \to \mathbb{R}$ satisfies $|f(x) - f(x')| \le c_k$ whenever the vectors x and x' differ only in the k-th coordinate. Let Z be the random variable $f(X_1, \dots, X_n)$. Then for any $t > 0$, $\mathbb{P}(|Z - \mathbb{E}Z| \ge t) \le 2 \exp\left(-2t^2 / \sum c_k^2\right)$.*

We prove the following lemma.

Lemma 9. *Let $r \ge 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n)$ satisfies $\sup_n \alpha(n) < \alpha^*_r$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. For every $\varepsilon > 0$ there is $\delta > 0$ such that, for all sufficiently large n,*

$$\mathbb{P}(F \text{ contains a full subformula of order} > \varepsilon n) < \exp(-\delta n).$$

Proof. Since the probability above is increasing in α, it is enough to prove the result for α constant, replacing $\alpha(n)$ by $\alpha = \sup_n \alpha(n)$. We will show that, with the required high probability, the pure literal rule leaves fewer than εn variables, establishing the lemma. (A full subformula is not affected by the pure literal rule, so if the "kernel" left is small, F contained no large subformula.)

Consider the following instantiation of the pure literal rule: Set $F_0 = F$, so $|F_0| = n$. For $i \ge 0$, obtain F_{i+1} from F_i by setting all pure literals to True, and then removing these literals and the

[1] An earlier version of the paper, [16], had an erroneous formula a factor of 2 smaller.

clauses they satisfied. Molloy showed that (for any $\alpha < \alpha^*$) there is a sequence $\lambda_s \to 0$ such that, for any s,

$$\mathbb{E}|F_s| = (1 + o(1))\lambda_s n.$$

Let us pick s such that $\lambda_s < \varepsilon/8$. The result will follow from a concentration argument which we now give in detail.

A *path of length* l in F is a sequence $v_0, C_0, v_1, C_1, \ldots, C_l, v_l$, alternating between variables and clauses, such that each clause C_i contains the variables that precede and follow it (either with or without negation). For a variable v and positive integer l, we define the ball $B_l(v)$ to be the subformula of F containing all variables and clauses that lie on paths of length at most l starting at v. (Note that each clause in a ball is fully supported by variables in it.)

It is part of Molloy's argument, and clear with a little thought, that the event that v belongs to F_s depends only on $B_s(v)$. We shall say that a variable v is *good* if it has the following two properties:

- v does not belong to $V(F_s)$ (the set of variables of F_s), and
- no variable in $B_s(v)$ belongs to more than $K(s, \alpha)$ clauses.

Here, $K(s, \alpha)$ is a constant chosen sufficiently large that the second property holds with probability at least $1 - \varepsilon/8$. There exists such a $K(s, \alpha)$ independent of n because the scaling of (1) was chosen precisely to make the local structure of an instance independent of n. For a simple rigorous argument, the degree of any variable in $B_s(v)$ is at most $|B_{s+1}(v)|$, $\mathbb{E}[|B_{s+1}(v)|]$ is obtained by multiplying the number of paths by their probability of being present and has an upper bound independent of n, and taking $K(s, \alpha)$ to be $8/\varepsilon$ times this value, the desired probability follows from Markov's inequality.

Since the first property occurs with probability $1 - \lambda_s + o(1)$, we see that for large enough n, v is good with probability greater than $1 - \varepsilon/4$. We will prove that, with failure probability $\exp(-\delta n)$, there are at least $(1-\varepsilon)n$ good variables. Now note that the pure literal rule can never set a variable belonging to a full subformula. Thus if H is a full subformula of F then $V(H) \subseteq \bigcap_{i=0}^{\infty} V(F_i)$. In particular, $V(H) \subset V(F_s)$ and so no good variable can belong to a full subformula. The claimed result is then immediate.

To prove our concentration bound, we first claim that changing a single clause in an instance cannot change the number of good variables by more than $2r^{s+1}K^s$. (This is the purpose of the second goodness condition.) Suppose we add a clause C to an instance I to obtain an instance I'. If adding C spoils a variable v (v is good in I but not in I'), C must contain some variable $u \in B_s(v)$. Choose a shortest path P from u to v. P has length at most s, and $P \subset I$ (it is shortest, so it doesn't contain C), thus $P \subset B_s(v)$, and since v was good in I, P contains no variables with degree (in I) more than K. Generating all paths of this sort, there are r choices for the variable $u \in C$, and from each variable at most K choices for the following clause and r choices for the succeeding variable, so there are at most $r^{s+1}K^s$ such paths, and at most that many spoiled variables. Therefore, adding a clause can decrease the number of good variables by at most $r^{s+1}K^s$, and similarly deleting a clause can create at most $r^{s+1}K^s$ good variables. The claim follows.

Finally, to use the Azuma-Hoeffding inequality (Lemma 8) we need to argue in terms of a fixed number of clauses. For this purpose we note that goodness is a monotonic property (if v is not good, adding clauses cannot make it good), and couple the original model $\mathcal{F}^r_{n,p}$ to one with a fixed and typically larger number of clauses. Specifically, first observe that the probability of being good is a continuous function of α (increasing α slightly adds a small linear number of new clauses, each of which spoils at most $r^{s+1}K^s$ good variables, a small fraction of the nearly n such variables). We can therefore choose $\alpha' > \alpha$ such that in an instance with clause probability $p' = \alpha' n^{-(r-1)}$, each variable is good with probability at least $1 - \varepsilon/3$. Let $p'' = (p + p')/2$ and $M = \lfloor p'' 2^r \binom{n}{r} \rfloor$. Define an *$M$-clause* model $\mathcal{F}^r_{n,M}$ where we sample M clauses uniformly *with replacement* from the set of all possible clauses, then discard duplicates (because of which this is not exactly the

analogue of the usual $G_{n,M}$ model). It is easy to check that, for some $\delta_0 > 0$, with probability $1 - O(\exp(-\delta_0 n))$, an instance of $\mathcal{F}^r_{n,p}$ has fewer clauses than one of $\mathcal{F}^r_{n,M}$ which in turn has fewer clauses than one of $\mathcal{F}^r_{n,p'}$. There is therefore a coupling between the three models in which, with probability $1 - O(\exp(-\delta_0 n))$, the corresponding random formulae satisfy $F_p \subset F_M \subset F_{p'}$.

We now complete the argument. By Lemma 8 (with $X_i = C_i$), in $\mathcal{F}^r_{n,M}$, with probability at least $1 - O(\exp(-\delta_1 n))$ the number of good variables is within $\varepsilon n/8$ of its expectation. By the coupling with $\mathcal{F}^r_{n,p'}$, this expectation is at least $(1 - \varepsilon/2)n$ (we inflate the $\varepsilon/3$ slightly to compensate for the exponentially small failure probability). So in $\mathcal{F}^r_{n,M}$, with exponentially small failure probability, we get at least $(1 - 2\varepsilon/3)n$ good variables. Finally, the coupling with $\mathcal{F}^r_{n,p}$ shows that, with exponentially small failure probability, we get at least $(1 - \varepsilon)n$ good variables. □

2.5. **Main results.** Consider the set of all MUFs. Order the set of values for excess as $\mathrm{ex}_1 < \mathrm{ex}_2 < \cdots$; by Lemma 1 these values are some subset of the positive integers). For $s > 0$, we write $\mathcal{F}_s$ for the set of MUFs F' with $\mathrm{ex}(F') = \mathrm{ex}_s$; note that by Lemma 1 each $\mathcal{F}_s$ is finite.

Theorem 10. *Fix $i > 0$. Let $r \geq 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = \Theta(1)$ satisfies $\sup_n \alpha(n) < \alpha^*_r$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. If we condition on the event that F is unsatisfiable and contains no MUF F' with $\mathrm{ex}(F') < \mathrm{ex}_i$ then, with probability $1 - O(1/n)$, the following statements hold:*

(i) F contains a unique MUF F_0.
(ii) $F_0 \in \mathcal{F}_i$.
(iii) For each $F' \in \mathcal{F}_i$, we have $\mathbb{P}(F_0 \cong F') \sim \frac{\alpha^{e(F')} 2^{|F'|}}{|\operatorname{aut} F'|}/Z$, where $Z = \sum_{F' \in \mathcal{F}_i} \frac{\alpha^{e(F')} 2^{|F'|}}{|\operatorname{aut} F'|}$.

Proof. This will follow by combining results from previous sections. Let C be the condition that F contain no MUF F' with $\mathrm{ex}(F') < \mathrm{ex}_i$ (but not that F is unsatisfiable).

Choose t_0 large enough and $\varepsilon_0 > 0$ small enough so that Corollary 7 applies with $s = \mathrm{ex}_i + 1$. Together with Corollary 9 (with $\varepsilon = \varepsilon_0$), we conclude that the probability that $F \in \mathcal{F}^r_{n,p}$ contains any full subformula on more than t_0 vertices is $o(n^{-s})$. This is also true after conditioning, since for any event E, $\mathbb{P}(E \mid C) = \mathbb{P}(E \wedge C)/\mathbb{P}(C) \leq \mathbb{P}(E)/\mathbb{P}(C) = (1 + O(1/n))\mathbb{P}(E)$.

There are finitely many possibilities for minimal unsatisfiable subformulae on t_0 or fewer vertices. From Lemma 5 and Lemma 3, for any F_0 with $\mathrm{ex}(F_0) \geq \mathrm{ex}_i$, $\mathbb{P}(F \supset F_0 \mid C) = (1 + O(1/n))\mathbb{P}(F \supset F_0) = (1 + O(1/n))\frac{2^{|F_0|}}{|\operatorname{aut} F_0|}\alpha^{e(F_0)} n^{-\mathrm{ex}(F_0)}$. When $F \in \mathcal{F}_i$, i.e., $\mathrm{ex}(F_0) = \mathrm{ex}_i$, this is a relatively likely event, with probability $\Theta(n^{-\mathrm{ex}_i})$; otherwise it is $O(1/n)$ less likely.

For any two MUFs F_1 and F_2 with $\mathrm{ex}(F_1), \mathrm{ex}(F_2) \geq \mathrm{ex}_i$, $\mathbb{P}(F$ contains non-nested copies of F_1 and $F_2 \mid C) = (1 + O(1/n))\mathbb{P}(F$ contains non-nested copies of F_1 and $F_2) = O(n^{-\mathrm{ex}_i + 1})$ by Lemma 4.

Now condition on the event that F is unsatisfiable, i.e., that at least one of the above cases occurs. Then the middle case, with $\mathrm{ex}(F_0) = \mathrm{ex}_i$, dominates the other cases. □

The same proof gives the analogous statement for minimal full subformulae. Consider the set of all MFFs, and order the set of values for excess as $\mathrm{ex}'_1 < \mathrm{ex}'_2 < \cdots$; again, these values are some subset of the positive integers. For $s > 0$, we write $\mathcal{F}'_s$ for the set of MFFs F' with $\mathrm{ex}(F') = \mathrm{ex}'_s$; note that by Lemma 1 each $\mathcal{F}'_s$ is finite.

Theorem 11. *Fix $i > 0$. Let $r \geq 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = \Theta(1)$ satisfies $\sup_n \alpha(n) < \alpha^*_r$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. If we condition on the event that F contains a full subformula, but no full subformula F' with $\mathrm{ex}(F') < \mathrm{ex}'_i$ then, with probability $1 - O(1/n)$, the following statements hold:*

(i) F contains a unique minimal full subformula F_0.
(ii) $F_0 \in \mathcal{F}'_i$.
(iii) For each $F' \in \mathcal{F}'_i$, we have $\mathbb{P}(F_0 \cong F') \sim \frac{\alpha^{e(F')} 2^{|F'|}}{|\operatorname{aut} F'|}/Z$, where $Z = \sum_{F' \in \mathcal{F}'_i} \frac{\alpha^{e(F')} 2^{|F'|}}{|\operatorname{aut} F'|}$.

We can also write an asymptotic expansion for the probability that F is unsatisfiable or that the pure literal rule fails (i.e., that F has a full subformula).

Theorem 12. *Let $r \geq 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = \Theta(1)$ satisfies $\sup_n \alpha(n) < \alpha_r^*$, and let $F \in \mathcal{F}_{n,p}^r$ be a random formula. For every full formula H there is a sequence of polynomials $p_1^{(H)}, p_2^{(H)}, \dots$ with rational coefficients such that, for any $s_{\max}$,*

$$\mathbb{P}(F \text{ contains a copy of } H) = \sum_{s=1}^{s_{\max}} p_s^{(H)}(\alpha) n^{-s} + O(n^{-s_{\max}-1}). \tag{9}$$

Furthermore, there is a sequence of polynomials $p_1, p_2, \dots$ with rational coefficients such that, for any $s_{\max}$ and any $\alpha < \alpha^*$,

$$\mathbb{P}(F \text{ is unsatisfiable}) = \sum_{s=1}^{s_{\max}} p_s(\alpha) n^{-s} + O(n^{-s_{\max}-1}), \tag{10}$$

and similarly a sequence $p_1', p_2', \dots$ such that

$$\mathbb{P}(\text{the pure literal rule fails on } F) = \sum_{s=1}^{s_{\max}} p_s'(\alpha) n^{-s} + O(n^{-s_{\max}-1}). \tag{10$'$}$$

Proof. Fix $s_{\max}$ and α. Note that (3) can be written as

$$\mathbb{E}X_H = \alpha^{e(H)} p_H(1/n), \tag{11}$$

where p_H is a polynomial of degree $\mathrm{ex}(H)$. The kth factorial moment of X_H is a sum of expectations $\mathbb{E}_{H'}$ over configurations H' consisting of the union of k distinct copies of H, and so is a sum of expressions like (11).

Now for $k \geq 1$, $\mathbb{P}(X_H = k)$ and $\mathbb{P}(X_H \geq k)$ can be written as alternating sums in the factorial moments (see [2, Section I.4]), and these sums satisfy the alternating inequalities. If K is fixed and sufficiently large then the Kth factorial moment has value $O(n^{-s_{\max}-1})$, as all its constituent configurations have excess larger than $s_{\max}$. Thus we can truncate our sum after a constant number of terms, with error $O(n^{-s_{\max}-1})$. Each term is of form (4), so we obtain an expression of form (9).

We obtain (10) similarly. Let $\mathcal{F}$ be the set of minimal unsatisfiable subformulae whose excess is at most $s_{\max}$, and let X be the number of subformulae of F that belong to $\mathcal{F}$. As in the previous case, asymptotic expansions for the factorial moments of X all have form (9), and once again applying inclusion-exclusion (and noting that we again have the alternating inequalities), truncating at the $n^{-s_{\max}}$ terms gives an asymptotic expansion of form (10). Minimal unsatisfiable subformulae of excess greater than $s_{\max}$ can be incorporated into the $O(n^{-s_{\max}-1})$ term by Lemmas 7 and 9. The argument for 10$'$ is identical, just phrased in terms of minimal full subformulae rather than minimal unsatisfiable subformulae. □

Let us note that it is only a finite (if tedious) computation to determine the polynomials p_s, and $p_s^{(H)}$ for any given H and s.

3. Structural results for sparse random graphs and hypergraphs

We now prove results on the k-core and k-colorability of a sparse random graph or hypergraph. The definitions, results, and proofs here precisely parallel those of Section 2.

We write $\mathcal{G}_r(n,p)$ for the random r-uniform hypergraph model analogous to $\mathcal{G}(n,p)$: a hypergraph $G \in \mathcal{G}_r(n,p)$ has vertex set $[n]$, and each possible edge (of size r) is independently present with probability p. We work with the scaling

$$p = \frac{cn}{\binom{n}{r}} = \alpha n^{-(r-1)}, \tag{12}$$

where p is the clause probability, cn is the expected number of clauses, and α is a convenient parametrization.

For an r-uniform hypergraph H we define

$$\mathrm{ex}(H) = (r-1)e(H) - |H|.$$

We say that H is *k-dense* if it has minimal degree $\delta(H) \geq k$. The hypergraph k-core is defined in the usual way, for example via the process detailed in the proof of Lemma 20, and it is k-dense. H is a minimal k-dense hypergraph if it is nonempty and has no proper k-dense subhypergraph.

Pittel, Spencer and Wormald [19] determined the threshold c_k for the appearance of a k-core in a random graph $G \in \mathcal{G}(n, c_k/n)$. They further showed that, for any fixed $c < c_k$ and $\varepsilon > 0$, the probability that $G \in \mathcal{G}(n, c/n)$ has a k-core of size bigger than εn is at most $\exp(-n^\delta)$ (in fact, they did rather more). Molloy [17] determined the k-core threshold $\alpha^{**} = \alpha^{**}_{k,r}$ for a random r-uniform hypergraph $G \in \mathcal{G}_r(n, \alpha n^{-(r-1)})$ and proved that for any fixed $\alpha < \alpha^{**}$ and $\varepsilon > 0$, the probability that $\mathcal{G}_r(n, \alpha n^{-(r-1)})$ has a k-core of size bigger than εn approaches 0.

Let us write $X_H(G)$ for the number of copies of H in G. Then

$$\begin{aligned} \mathbb{E}X_H &= \frac{1}{|\operatorname{aut} H|}\binom{n}{|H|}|H|!p^{e(H)} \\ &= (1+O(1/n))\tfrac{1}{|\operatorname{aut} H|}\alpha^{e(H)}n^{-\mathrm{ex}(H)}. \end{aligned} \tag{13}$$

Lemma 13. *Suppose that $r, k \geq 2$ and $r + k > 4$. If H is a k-dense, r-uniform hypergraph then*

$$\mathrm{ex}(H) \geq \frac{(k-1)(r-1)-1}{r}|H|.$$

Furthermore, for every $s > 0$, there are (up to isomorphism) only finitely many k-dense graphs H with $\mathrm{ex}(H) = s$.

Proof. If $\delta(H) \geq k$ then $e(H) \geq k|H|/r$ and so

$$\mathrm{ex}(H) \geq k|H|(r-1)/r - |H| = \frac{(k-1)(r-1)-1}{r}|H|.$$

So if $\mathrm{ex}(H) = s$ then $|H| \leq rs/[(k-1)(r-1)-1]$, which implies that there are only finitely many possibilities for H. □

Note that the k-core is necessarily k-dense. It follows that there are only finitely many possible k-cores of each excess.

Proposition 14. *Suppose that $r, k \geq 2$ and $r+k > 4$. For k-dense, r-uniform hypergraphs H_1 and H_2, with $H_1 \not\subseteq H_2$, $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + 1$.*

Proof. If $V(H_1) \subseteq V(H_2)$ then $|H_1 \cup H_2| = |H_2|$ while $e(H_1 \cup H_2) > e(H_2)$ implying $\mathrm{ex}(H_1 \cup H_2) > \mathrm{ex}(H_2)$ which by integrality means $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + 1$.

Otherwise, let $t = |V(H_1) \setminus V(H_2)| > 0$. Then $H_1 \cup H_2$ contains at least kt/r more edges than H_2 (since each vertex in $V(H_1) \setminus V(H_2)$ is incident with at least k edges). So $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + kt(r-1)/r - t > \mathrm{ex}(H_2)$. Since ex is integer-valued, this implies $\mathrm{ex}(H_1 \cup H_2) \geq \mathrm{ex}(H_2) + 1$. □

Claim 15. *Let $r, k \geq 2$, $r + k > 4$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $G \in \mathcal{G}_r(n, p)$ be a random hypergraph. For any fixed k-dense, r-uniform hypergraph H,*

$$\mathbb{P}(\exists \text{ a copy of } H \text{ in } G) = (1+O(1/n))\tfrac{1}{|\operatorname{aut} H|}\alpha^{e(H)}n^{-\mathrm{ex}(H)}.$$

Proof. With X_H the number of copies of H in G, we have from (13) that

$$\mathbb{E}X_H = (1+O(1/n))\tfrac{1}{|\operatorname{aut} H|}\alpha^{e(H)}n^{-\mathrm{ex}(H)},$$

while

$$\mathbb{E}[X_H(X_H-1)] = O(1)\ \alpha^{e(H)+1} n^{-(\mathrm{ex}(H)+1)} = O(\alpha/n)\ \mathbb{E}[X_H]$$

and the rest of the proof follows as for Claim 3. □

Lemma 16. *Let $r, k \geq 2$, $r+k > 4$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $G \in \mathcal{G}_r(n,p)$ be a random hypergraph. Let H_1 and H_2 be fixed k-dense, r-uniform hypergraphs. Then*

$$\begin{aligned}&\mathbb{P}(G \text{ contains non-nested copies of } H_1 \text{ and } H_2)\\ &\qquad = O(n^{-\max\{\mathrm{ex}(H_1),\mathrm{ex}(H_2)\}-1}) \quad .\end{aligned}$$

Proof. Another proof without changes. □

Lemma 17. *Let $r, k \geq 2$, $r+k > 4$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $G \in \mathcal{G}_r(n,p)$ be a random hypergraph. If $H_1, \ldots, H_s$ are distinct minimal k-dense, r-uniform hypergraphs (or minimal non-k-colorable r-uniform hypergraphs) then*

$$\mathbb{P}(G \supset H_1 \mid G \not\supset H_2, \ldots, G \not\supset H_s) = (1 + O(1/n))\mathbb{P}(G \supset H_1).$$

Proof. Another proof without changes. □

Lemma 18. *Let $r, k \geq 2$, $r+k > 4$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n)$, and let $G \in \mathcal{G}_r(n,p)$ be a random hypergraph. For $1 \leq t \leq n/\alpha^{1/(r-1)}$, the probability that G contains any k-dense subhypergraph with t variables is at most*

$$\left(\left(e\alpha^{k/r}\right)(t/n)^{k-1-1/r}\right)^t. \tag{14}$$

Proof. We modify the proof of Lemma 6. Order the vertices as $v_1 < v_2 < \cdots$. A k-dense subhypergraph H of G with order t must contain at least kt/r edges. We let $s = \lceil kt/r \rceil$ and define a subhypergraph H^* of H with s edges as follows. Let L be the set of t vertices occurring in edges of H. Let x_1 be the smallest vertex in L, and let C_1 be the lexicographically smallest edge of H (sorting the vertices within each edge as above) that contains x_1. For $i = 2, \ldots, s$, let x_i be the smallest vertex in L that is not covered k times by C_j, $j < i$, and let C_i be the lexicographically smallest edge of H that contains x_i. (This is well defined since we are always excluding at most $s-1$ edges, which together contain at most $(s-1)r < kt$ vertex occurrences.) We then take H^* to be the edge set $C_1, \ldots, C_s$.

The number of hypergraphs of type H^* that could possibly be subhypergraphs of G is at most $\binom{n}{t}\binom{t}{r-1}^s$. Let X be the number of k-dense subhypergraphs of G with order t, and let Y be the number of subhypergraphs of type H^* of G. Then $X > 0$ implies $Y > 0$, so

$$\begin{aligned}\mathbb{P}(X>0) \leq \mathbb{P}(Y>0) \leq \mathbb{E}(Y) &\leq \binom{n}{t}\binom{t}{r-1}^s p^s\\ &\leq (en/t)^t (t)^{s(r-1)} (\alpha/n^{r-1})^s\\ &= (en/t)^t \left(\alpha(t/n)^{(r-1)}\right)^s\\ &\leq (en/t)^t \left(\alpha(t/n)^{(r-1)}\right)^{kt/r},\end{aligned}$$

which equals (14). □

Corollary 19. *Let $r, k \geq 2$, $r+k > 4$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = O(1)$, and let $G \in \mathcal{G}_r(n,p)$ be a random hypergraph. For any positive integer s, there exist an integer $t_0 > 0$ and a real value $\varepsilon_0 > 0$ such that the probability that G contains a k-dense subhypergraph with between t_0 and $\varepsilon_0 n$ vertices is $o(n^{-s})$.*

Proof. As before. □

Recall that we defined $\alpha^{**}_{k,r}$ to be the k-core threshold for $\mathcal{G}_r(n, \alpha n^{-(r-1)})$.

Lemma 20. *Let $r, k \geq 2$, $r + k > 4$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n)$ satisfies $\sup_n \alpha(n) < \alpha^{**}_{k,r}$, and let $G \in \mathcal{G}_r(n, p)$ be a random hypergraph. For every $\varepsilon > 0$ there is $\delta > 0$ such that, for all sufficiently large n,*

$$\mathbb{P}(G \text{ contains a } k\text{-dense subhypergraph with order} > \varepsilon n) < \exp(-\delta n).$$

Proof. We follow the argument of Lemma 9. We use the following process for generating the k-core: Set $G_0 = G$, so $|G_0| = n$. For $i \geq 0$, obtain G_{i+1} from G_i by deleting (in a single round) all vertices of degree at most $k - 1$, and all edges incident on any such vertex. The k-core is $G_\infty = G_n$. As with satisfiability, Molloy showed that (for $\alpha < \alpha^{**}$) there is a sequence $\lambda_s \to 0$ such that, for any s,

$$\mathbb{E}|G_s| = (1 + o(1))\lambda_s n.$$

A ball $B_s(v)$ has the usual hypergraph definition analogous to the ball definition in the proof of Lemma 9, and each edge in a ball is fully supported by vertices in it. We shall say that a vertex v is *good* if it has the following two properties:

- v does not belong to $V(G_s)$, and
- no vertex in $B_s(v)$ has degree more than $K(s, \alpha)$.

The rest of the proof is as before. □

Consider the set of all minimal non-k-colorable hypergraphs, order the set of values for excess as $\mathrm{ex}_1 < \mathrm{ex}_2 < \cdots$, and let $\mathcal{G}_i$ be the set of non-k-colorable hypergraphs with excess i. Similarly, let the minimal k-dense hypergraphs have excesses $\mathrm{ex}'_1 < \mathrm{ex}'_2 < \cdots$ and let $\mathcal{G}'_i$ be the set of minimal k-dense hypergraphs with excess i. Then we have the analogues of Theorems 10, 11, and 12, by the same reasoning.

Theorem 21. *Fix $i > 0$. Let $r, k \geq 2$, $r + k > 4$, let $p = \alpha n^{-(r-1)} = \Theta(1)$ where $\alpha = \alpha(n)$ satisfies $\sup_n \alpha(n) < \alpha^{**}_{k,r}$, and let $G \in \mathcal{G}_r(n, p)$ be a random hypergraph. If we condition on the event that G is non-k-colorable and contains no minimal non-k-colorable G' with $\mathrm{ex}(G') < \mathrm{ex}_i$ then, with probability $1 - O(1/n)$, the following statements hold:*

(i) G contains a unique minimal non-k-colorable G_0.
(ii) $G_0 \in \mathcal{G}_i$.
(iii) For each $G' \in \mathcal{G}_i$, we have $\mathbb{P}(G_0 \cong G') \sim \frac{\alpha^{e(G')}}{|\operatorname{aut} G'|}/Z$, where $Z = \sum_{G' \in \mathcal{G}_i} \frac{\alpha^{e(G')}}{|\operatorname{aut} G'|}$.

Theorem 22. *Fix $i > 0$. Let $r, k \geq 2$, $r + k > 4$, let $p = \alpha n^{-(r-1)} = \Theta(1)$ where $\alpha = \alpha(n)$ satisfies $\sup_n \alpha(n) < \alpha^{**}_{k,r}$, and let $G \in \mathcal{G}_r(n, p)$ be a random hypergraph. If we condition on the event that G contains a nonempty k-core, but no nonempty k-core G' with $\mathrm{ex}(G') < \mathrm{ex}'_i$ then, with probability $1 - O(1/n)$, the following statements hold:*

(i) G contains a unique minimal nonempty k-core G_0.
(ii) $G_0 \in \mathcal{G}'_i$.
(iii) For each $G' \in \mathcal{G}'_i$, we have $\mathbb{P}(G_0 \cong G') \sim \frac{\alpha^{e(G')}}{|\operatorname{aut} G'|}/Z$, where $Z = \sum_{G' \in \mathcal{G}'_i} \frac{\alpha^{e(G')}}{|\operatorname{aut} G'|}$.

Theorem 23. *Let $r \geq 3$, let $p = \alpha n^{-(r-1)}$ where $\alpha = \alpha(n) = \Theta(1)$ satisfies $\sup_n \alpha(n) < \alpha^*_r$, and let $F \in \mathcal{F}^r_{n,p}$ be a random formula. For every k-dense hypergraph H there is a sequence of polynomials $p^{(H)}_1, p^{(H)}_2, \ldots$ with rational coefficients such that, for any $s_{\max}$,*

$$\mathbb{P}(G \text{ contains a copy of } H) = \sum_{s=1}^{s_{\max}} p^{(H)}_s(\alpha) n^{-s} + O(n^{-s_{\max}-1}). \tag{15}$$

Furthermore, there is a sequence of polynomials $p_1, p_2, \ldots$ with rational coefficients such that, for any $s_{\max}$ and any $\alpha < \alpha^*$,

$$\mathbb{P}(G \text{ is non-}k\text{-colorable}) = \sum_{s=1}^{s_{\max}} p_s(\alpha) n^{-s} + O(n^{-s_{\max}-1}), \tag{16}$$

and similarly a sequence $p'_1, p'_2, \ldots$ such that

$$\mathbb{P}(G \text{ has a nonempty } k\text{-core}) = \sum_{s=1}^{s_{\max}} p'_s(\alpha) n^{-s} + O(n^{-s_{\max}-1}). \tag{16'}$$

4. Conclusion

4.1. **Examples.** For graphs, i.e., hypergraphs with $r = 2$, any k-dense graph on n vertices has $n \geq k+1$ (each degree is at least k) and at least $kn/2$ edges, thus has excess at least $(k/2-1)n$; this is uniquely minimized by $n = k+1$ and the graph K_{k+1}, with excess $(k-2)(k+1)/2$, $k(k+1)/2$ edges, and $|\operatorname{aut} K_{k+1}| = (k+1)!$. Since K_{k+1} is non-k-colorable, it is also the unique smallest non-k-colorable graph. Thus for $k \geq 3$, $\alpha < \alpha^{**}_{k,2}$, and $G \in \mathcal{G}(n, \alpha/n)$,

$$\mathbb{P}(G \text{ is not } k\text{-colorable}) = (1+O(1/n))\tfrac{1}{(k+1)!}\, \alpha^{k(k+1)/2} n^{-(k-2)(k+1)/2}, \text{ and}$$
$$\mathbb{P}(G \text{ has a nonempty } k\text{-core}) = (1+O(1/n))\tfrac{1}{(k+1)!}\, \alpha^{k(k+1)/2} n^{-(k-2)(k+1)/2}.$$

Furthermore, if G has nonempty k-core then with probability $1+O(1/n)$ its k-core is a single copy of K_{k+1}; the same conclusion follows if G is not k-colorable.

For random r-SAT formulae, any full formula on t variables has excess at least $(r-2)t/r$, and this is minimized uniquely by $t = r$ and the formula F_L consisting of the 2 clauses $(X_1, \ldots, X_r)$ and $(\bar{X}_1, \ldots, \bar{X}_r)$, with excess $r-2$ and $2 \cdot r!$ automorphisms. Thus, for $r \geq 3$ and $F \in \mathcal{F}^r_{n,p}$,

$$\mathbb{P}(\text{the pure literal rule fails to satisfy } F) = (1+O(1/n))\tfrac{1}{2 \cdot r!} \alpha^2 n^{-(r-2)}. \tag{17}$$

Furthermore, if the pure literal rule fails to satisfy F then with probability $1+O(1/n)$ its pure literal core is a single copy of F_L, which is satisfiable, in contrast to the graph case, where we have seen that the k-core is almost surely the non-k-colorable graph K_{k+1}.

4.2. **2-SAT and 2-CSP.** For random formulas, we have assumed throughout that $r \geq 3$, because this is needed for Lemma 1. Also, we leave unresolved what happens between the pure literal and satisfiability thresholds. However, much is already known about random 2-SAT, and in this case the thresholds are equal, both having $\alpha = 1/2$. Chvátal and Reed [4] show that a 2-SAT formula is unsatisfiable iff it contains a "bicycle", and it is straightforward to compute the likelihoods of bicycles of various sizes. Our earlier paper [20] exploited the typically small size of the 2-core of a random graph $G \in \mathcal{G}(n, \alpha/n)$ with $\alpha < 1$ (a threshold above which the core jumps to linear size) to give an algorithm running in expected time $O(n)$ for "random" instances of any Max 2-CSP below this threshold; the class of optimization problems Max 2-CSP includes Max 2-Sat and Max Cut.

4.3. **Very sparse instances.** For very sparse instances ($\alpha \to 0$ very quickly), our results need a little modification, as the preference order for small subinstances must be changed. For instance if $p = n^{-\log n}$ then FFs will appear primarily in order of the number of clauses and only secondarily in terms of number of variables (rather than in terms of excess).

4.4. **Structural results.** Our results on small subformulae hold for any constant density. However, above some threshold large subformulae appear. Our structure theory for random unsatisfiable formulas applies below the pure literal threshold, because we know there are no large full subformulas in this range. From the other side, we know by Chvátal-Szemerédi [5] (or from our analysis) that large unsatisfiable subformulae (therefore large full formulae) appear above the satisfiability threshold. (At a density $\alpha = \Theta(1)$ any constant above the satisfiability threshold, an instance is **whp** unsatisfiable [10], but our results for subformulae of small and medium size apply for any $\alpha = \Theta(1)$, so full [and potentially unsatisfiable] subformulae of up to small linear size occur with small probability, thus the obstruction to satisfiability must **whp** be a large minimal unsatisfiable subformula.) It would be most interesting to know what happens for formulas between the pure literal and satisfiability thresholds.

Specifically, are large minimal unsatisfiable subformulae unlikely between the two thresholds, as are large full subformulae below the pure literal threshold? Concretely, let $c_r(n)n^{-(r-1)}$ be a threshold function for r-SAT; recall that $c_r(n)$ is believed but not known to converge to a constant.

Question 1. *Let $r \geq 3$, $\varepsilon > 0$, and $p = \alpha n^{-(r-1)}$, with $\alpha(n) \leq (1-\varepsilon)c_r(n)$ for all n. Let $q(n)$ be the probability that a random formula $F \in \mathcal{F}^r_{n,p}$ contains a minimal unsatisfiable subformula on at least εn vertices. Is $q(n) = n^{-\omega(1)}$?*

A positive answer would immediately translate into a proof of a structural theorem.

4.5. **Algorithms.** The behavior of algorithms up to the satisfiability threshold is unclear. However, it is easy to give algorithms for sufficiently sparse instances. For instance:

Theorem 24. *For all r, for all sufficiently small α there is an expected polynomial-time algorithm to decide the satisfiability of a random formula $F \in \mathcal{F}^r_{n,p}$, outputting an assignment satisfying as many clauses as possible and (if F is unsatisfiable) a minimal unsatisfiable subformula.*

Proof. This follows from (7), for some α smaller (likely much smaller) than the pure literal threshold α^*.

We first apply the pure literal rule, taking time $O^*(1)$ (a notation that hides factors polynomial in the input parameters) and leaving a full subformula on t variables (if $t = 0$, F is satisfied and we are done). If there are t remaining variables, we now try all 2^t possible assignments, taking time $O^*(2^t)$. If α is sufficiently small, then (7) is at most 4^{-t} for all $t \geq 1$, and the expected running time is at most $\sum_{t\geq 1} O^*(1)2^t 4^{-t} = O^*(1)$.

To produce a minimal unsatisfiable subformula, or list all such subformulas, again we apply pure literal until we are left with a full subformula with t variables and s clauses. Note that there are at most $\binom{n}{t}\binom{2^r\binom{t}{r}}{s}$ such formulae, each of which is present with probability at most p^s, with $s \geq 2t/r$. We now look at all 2^s subformulae, and for each we check all 2^t assignments of our remaining variables (we can easily order the subformulae so that we can search for a minimal unsatisfiable subformula). This takes expected time at most

$$\begin{aligned}
\sum_{t\geq 1}\sum_{s\geq 2t/r} 2^t 2^s \binom{n}{t}\binom{2^r\binom{t}{r}}{s} p^s &\leq \sum_{t\geq 1}\sum_{s\geq 2t/r} \left(\frac{2en}{t}\right)^t \left(\frac{2^{r+1}et^r}{s}\right)^s \left(\frac{\alpha}{n^{r-1}}\right)^s \\
&\leq \sum_{t\geq 1}\sum_{s\geq 2t/r} (2e)^{(r+1)s+t}\alpha^s \left(\frac{n}{t}\right)^t \left(\frac{t^r}{2t/r}\right)^s \left(\frac{1}{n^{r-1}}\right)^s \\
&\leq \sum_{t\geq 1}\sum_{s\geq 2t/r} (2er)^{2rs}\alpha^s \left(\frac{t}{n}\right)^{(r-1)s-t} \\
&\leq \sum_{t\geq 1}\sum_{s\geq 2t/r} (2er)^{2rs}\alpha^s
\end{aligned}$$

$$\leq \sum_{t \geq 1} 2^{-t}$$
$$\leq 1,$$

provided α is small enough. Since the initial application of pure literal takes time $O^*(1)$ we are done. □

If the structural results extend up to the satisfiability threshold, then *most* unsatisfiable instances in the satisfiable regime have a small witness, and so can be identified quickly. This would affirmatively answer the following question.

Question 2. *Suppose $\varepsilon > 0$ and $\alpha = \alpha(n) \leq (1-\varepsilon)c_r(n)$ for all n. Is there a polynomial-time algorithm that, **whp**, proves unsatisfiability for a random unsatisfiable formula $F \in \mathcal{F}_r(n, \alpha n^{-(r-1)})$?*

More ambitiously, we could hope for algorithms that succeed *always*, and run in polynomial expected time (possibly only for smaller densities α).

Question 3. *Suppose $\varepsilon > 0$ and $\alpha = \alpha(n) \leq (1-\varepsilon)c_r(n)$ for all n. Is there an algorithm that, for a random unsatisfiable formula $F \in \mathcal{F}_r(n, \alpha n^{-(r-1)})$ proves unsatisfiability in polynomial expected time?*

4.6. **Graphs and hypergraphs.** In the graph and hypergraph context, we would like to know what happens between the k-core threshold $\alpha^{**}_{k,r}$ and a k-colorability threshold $d_{k,r}(n)n^{-(r-1)}$, recalling that $d_{k,r}(n)$ is believed but not known to converge to a constant. Here the essential question is the analogue of Question 1: are large minimal non-k-colorable subhypergraphs unlikely between the two thresholds (as large k-dense subhypergraphs are below k-core threshold)?

A result like Theorem 24 can easily be proved for hypergraph coloring (see also [6] for results on coloring sparse random graphs). With $r, k \geq 2$, $r + k > 4$, $\varepsilon > 0$, $p = \alpha n^{-(r-1)}$, and $\alpha(n) \leq (1-\varepsilon)d_{k,r}(n)$, there are also the obvious analogues of Questions 2 and 3: are there are algorithms that are efficient (almost always, or in expectation) for k-coloring random r-uniform hypergraphs below the k-coloring threshold?

References

[1] D. Achlioptas and Y. Peres, The threshold for random k-SAT is $2^k \log 2 - O(k)$, *Journal of the American Mathematical Society* **17** (2004), 947–973

[2] B. Bollobás, Random Graphs, Academic Press, 1985

[3] A. Broder, A. Frieze and E. Upfal, On the satisfiability and maximum satisfiability of random 3-CNF formulas, Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (Austin, TX, 1993), 322–330, ACM, New York, 1993

[4] V. Chvátal and B. Reed, *Mick gets some (the odds are on his side)*, 33th Annual Symposium on Foundations of Computer Science (Pittsburgh, PA, 1992), IEEE Comput. Soc. Press, Los Alamitos, CA, 1992, pp. 620–627

[5] V. Chvátal and E. Szemerédi, Many hard examples for resolution, *J. Assoc. Comput. Mach.* **35** (1988), 759–768

[6] A. Coja-Oghlan and A. Taraz, Exact and approximative algorithms for coloring $G(n,p)$, *Random Structures and Algorithms* **24** (2004), 259–278

[7] A. Coja-Oghlan, *A better algorithm for random k-SAT*, SIAM Journal on Computing **39** (2010), 2823–2864.

[8] A. Coja-Oghlan, M. Krivelevich, and D. Vilenchik, *Why almost all satisfiable k-CNF formulas are easy*, 2007 Conference on Analysis of Algorithms, AofA 07, Discrete Math. Theor. Comput. Sci. Proc., AH, Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2007, pp. 89–101.

[9] P. Erdős, D.J. Kleitman, and B.L. Rothschild, Asymptotic enumeration of K_n-free graphs, In International Colloquium on Combinatorial Theory (Rome, 1973), Vol. 2, 19–27. Atti dei Convegni Lincei, No. 17, *Accad. naz. Lincei, Rome*, 1976.

[10] E. Friedgut, Necessary and sufficient conditions for sharp thresholds of graph properties, and the k-SAT problem, *J. Amer. Math. Soc.* **12** (1999), 1017–1054

[11] M. Luby, M. Mitzenmacher and M. Shokrollahi, Analysis of random processes via And-Or tree evaluation, Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998), 364–373, ACM, New York, 1998

[12] C. McDiarmid, On the method of bounded differences, *in* Surveys in combinatorics, 1989 (Norwich, 1989), London Math. Soc. Lecture Note Ser. **141**, Cambridge Univ. Press 1989, pp. 148–188
[13] M. Mitzenmacher, *Tight thresholds for the pure literal rule*, Technical Note 1997-011, Digital Systems Research Center, Palo Alto (1997)
[14] M. Molloy, A Gap Between the Appearance of a k-Core and a (k+1)-Chromatic Graph, *Random Structures and Algorithms* **8** (1996), 159–160
[15] M. Molloy, Thresholds for colourability and satisfiability in random graphs and Boolean formulae, Surveys in combinatorics, 2001 (Sussex), 165–197, London Math. Soc. Lecture Note Ser., 288, Cambridge Univ. Press, Cambridge, 2001
[16] M. Molloy, The pure literal rule threshold and cores in random hypergraphs, Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA, New Orleans, Louisiana, 2004), 2004, 672–681, SIAM, Philadelphia, PA, USA
[17] M. Molloy, Cores in random hypergraphs and Boolean formulas, *Random Structures Algorithms* **27** (2005), 124–135
[18] H.J. Prömel, T. Schickinger, A. Steger, On the structure of clique-free graphs, *Random Structures Algorithms* **19** (2001), no. 1, 37–53.
[19] B. Pittel, J. Spencer and N. Wormald, Sudden emergence of a giant k-core in a random graph, *J. Combin. Theory Ser. B* **67** (1996), 111–151
[20] A.D. Scott and G.B. Sorkin, Solving sparse random instances of Max Cut and Max 2-CSP in linear expected time, *Combinatorics, Probability and Computing* **15** (2006), 281–315

(Alexander D. Scott) Mathematical Institute, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK

E-mail address: scott@maths.ox.ac.uk

(Gregory B. Sorkin) Department of Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights NY 10598, USA

E-mail address: sorkin@watson.ibm.com